DOCUMENT RESUME

ED 193 308 TM 800 623

AUTHOR Mellenbergh, Gideon J.: Vijn, Pieter
TITLE Mental Test Data and Contingency Tables.
PUB DATE Jun 80
NOTE 15p.: Paper presented at the 1980 European Meeting of the Psychometric Society (Groningen, The Netherlands, June 19-21, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Black Students: *Cognitive Measurement: Foreign Countries: *Latent Trait Theory: *Mathematical Models: Maximum Likelihood Statistics: Multidimensional Scaling: *Multivariate Analysis: Psychometrics: Sampling: *Test Bias: Test Theory: White Students
IDENTIFIERS Frequency Analysis: Frequency Data: *Log Linear Models: Netherlands: *Rasch Model

ABSTRACT

Data are summarized in Scheuneman's Score x Group x Response frequency table in order to investigate item bias. The data can arise from two different sampling models: (1) multinomial sampling in which a fixed sample size is used and the responses are cross-classified according to score, group, and response: and (2) product-multinomial sampling in which for each group a fixed sample size is used and the responses are cross-classified according to score and response: Data for both sampling models were analyzed using two logit models, i.e. special cases of log linear models, and results of the procedure were applied to Scheuneman's data using the program ECTA. The item was uniformly biased as shown by whites performing better than blacks in all score categories. Using a frequency table derived from Table 2 of Perline, Wright, and Wainer's nine-item scale for parole decision data, the linear logit model and the Rasch Model were found to be equivalent. Consequently, the estimates for the parameters in the log linear model yield unconditional maximum likelihood estimates for the parameters in the Rasch Model. (RL)

# Mental Test Data and Contingency Tables

Gideon J. Mellenbergh and Pieter Vijn



Paper presented at the 1980 European Meeting of the Psychometric Society, Groningen, The Netherlands, 19-21 June 1980.

Mental tests usually consist of items or stimuli with discrete response categories such as correct and incorrect answers, true or false, strenght of agreement in a number of categories and so on. In classical test theory the item responses are scored and the item scores are combined in a total test score. For example, in achievement tests the items are usually scored zero or one and the total score equals the number correctly answered items. In latent trait theory, however, the latent scale is constructed taking account of the discrete nature of the responses.

Recently general methods for the analysis of multi-dimensional tables have been developed (Bishop, Fienberg & Holland, 1975). It is hypothesized that many psychometric models and methods can be translated into contingency table methods. In this paper two topics are discussed. First, the assessment of item bias from a contingency table and second, the formulation of the Rasch model as a loglinear model.

## Assessing Item Bias

Scheuneman (1979) defined an unbiased item as one "for which the probability of a correct response is the same for all persons of a given ability level". To investigate item bias the data are summarized in a Score x Group x Response frequency table. Constructing the table scores with small number of subjects must be combined. As an example consider the frequencies in the three-dimensional table derived from Scheuneman's Table 1:

Table 1

Frequencies in Score x Group x Response Table, Scheuneman's Data

| | Response (m) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct (m=1) | | | Incorrect (m=0) | | | Total | | Total |
| Group (j) | Black (j=1) | White (j=2) | Total | Black (j=1) | White (j=2) | Total | Black (j=1) | White (j=2) | Subjects |
| Score (i) | | | | | | | | | |
| 13-14 (i=1) | 22 | 300 | 322 | 3 | 15 | 18 | 25 | 315 | 340 |
| 12 (i=2) | 18 | 99 | 117 | 6 | 11 | 17 | 24 | 110 | 134 |
| 10-11 (i=3) | 23 | 93 | 116 | 25 | 25 | 50 | 48 | 118 | 166 |
| 1-9 (i=4) | 14 | 33 | 47 | 51 | 59 | 110 | 65 | 92 | 157 |

These data can arise from two different sampling models. First, multinomial sampling: A fixed sample size is used and the responses are cross-classified according to Score, Group, and Response. Second, product-multinomial sampling: For each group a fixed sample size is used and the responses are cross-classified according to Score and Response (Fienberg, 1977, section 2.4). For both sampling model the data can be analyzed with loglinear models.

For dichotomously scored items the Response variable has only two categories, i.e. a correct and an incorrect response. The response ratio is the ratio of the frequencies of the correct and the incorrect responses. For example, the estimated response ratio for Blacks in Score category 10-11 in Table 1 is 23/25= .92. For the special case of only two response categories loglinear models can be transformed to logit models for the response ratio (Fienberg, 1977, section 6.2). In this paper logit models are described to assess item bias and to investigate the nature of the bias.

A linear logit model for the response ratio in the $i$th Score category ($i = 1,2,\ldots, r$) and $j$th Group ($j = 1,2,\ldots, k$) is:

$$\ln(F_{ij1}/F_{ij0}) = C + S_i + G_j, \qquad (1)$$

with constraints:

$$\sum_{i=1}^{r} S_i = 0, \qquad (2)$$

$$\sum_{j=1}^{k} G_j = 0. \qquad (3)$$

$F_{ij1}$ and $F_{ij0}$ are the population frequencies of the correct and incorrect response in the $i$th Score category and $j$th Group; ln means the natural logarithm. The parameter $C$ is a constant. The parameters $S_i$ and $G_j$ can be interpreted similarly to the main effect parameters in an analysis of variance model: $S_i$ is the effect of the $i$th Score category and $G_j$ the effect of the $j$th Group on the response ratio.

If this model does not fit the data a parameter for the interaction of the Score categories and the Group is needed. For example, if the response ratio in the low Score categories is higher for Blacks than for Whites whereas the response ratio in the high Score categories is lower for Blacks than for Whites, Score and Group interact and the item is biased. The nature of the bias is rather complicated. An item for which the interaction between Score and Group is necessary to explain the response ratios will be called not-uniformly biased.

If the model fits the data a second logit model is considered:

$$\ln(\underline{F}_{ij1}/\underline{F}_{ij0}) = \underline{C} + \underline{S}_i, \qquad (4)$$

with constraint Formula 2. If this model does not fit the data the item is also biased. This conclusion is strengthened by a significant difference in fit of both models. In this case the bias is uniform: For all Score categories the difference between the response ratios of Blacks and Whites is constant. In this case the item will be called uniformly biased. Finally, if model Formula 4 fits the data the item is not biased.

Logit models are special cases of loglinear models. The loglinear model corresponding to the logit model Formula 1 is:

$$\ln \underline{F}_{ijm} = \underline{u} + \underline{u}_{1(i)} + \underline{u}_{2(j)} + \underline{u}_{3(m)} + \underline{u}_{12(ij)} + \underline{u}_{13(im)} + \underline{u}_{23(jm)}, \qquad (5)$$

with constraints

$$\sum_{i=1}^{r} \underline{u}_{1(i)} = \sum_{j=1}^{k} \underline{u}_{2(j)} = \sum_{m=0}^{1} \underline{u}_{3(m)} = 0, \qquad (6)$$

$$\sum_{i=1}^{r} \underline{u}_{12(ij)} = \sum_{j=1}^{k} \underline{u}_{12(ij)} = \sum_{i=1}^{r} \underline{u}_{13(im)} = \sum_{m=0}^{1} \underline{u}_{13(im)} =$$

$$\sum_{j=1}^{k} \underline{u}_{23(jm)} = \sum_{m=0}^{1} \underline{u}_{23(jm)} = 0. \qquad (7)$$

From these Formulas follows the linear logit model:

$$\ln(F_{ij1}/F_{ij0}) = (u_{3(1)} - u_{3(0)}) + (u_{13(i1)} - u_{13(i0)}) +$$
$$(u_{23(j1)} - u_{23(j0)}) = 2u_{3(1)} + 2u_{13(i1)} + 2u_{23(j1)} =$$
$$C + S_i + G_j. \tag{8}$$

In the same way setting the parameters $u_{23(jm)}$ in model Formula 5 equal to zero yields the logit model Formula 4. Consequently, procedures used for loglinear models can be applied to logit models.

For assessing the fit of a model Pearson's chi-square,

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \sum_{m=0}^{1} (f_{ijm} - \hat{f}_{ijm})^2 / \hat{f}_{ijm}, \tag{9}$$

and the likelihood ratio chi-square

$$G^2 = 2 \sum_{i=1}^{r} \sum_{j=1}^{k} \sum_{m=0}^{1} f_{ijm} \ln(f_{ijm} / \hat{f}_{ijm}) \tag{10}$$

are used. In these Formulas $f_{ijm}$ is the frequency obtained in a sample under the multinomial or product-multinomial sampling model. The $\hat{f}_{ijk}$ is the estimate of the model expected frequency. Both statistics are asymptotically chi-square distributed. The degrees of freedom for model Formula 4 are $(k-1)r$ and for model Formula 1 $(k-1)(r-1)$ (Fienberg, 1977, p. 37). A property of the $G^2$-statistic is that the difference in $G^2$ values of the nested models Formula 4 and 1 is asymptotically chi-square distributed with $(k-1)r-(k-1)(r-1) = (k-1)$ degrees of freedom. The computations can be done with computer programs such as ECTA (Goodman & Fay, 1974) and BMDP3F (Dixon & Brown, 1977).

The results of the procedure applied to Scheuneman"s data using the program ECTA are reported in Table 2.

Table 2

Models Fitted to Scheuneman's Data

| Model | $X^2$ | $G^2$ | Df | Critical Chi-square 5% | 1% |
|---|---|---|---|---|---|
| Score and Group Effect (Formula 1) | 1.74 | 1.72 | 3 | 7.81 | 11.34 |
| Score Effect (Formula 4) | 25.63 | 23.93 | 4 | 9.49 | 13.28 |
| Difference $G^2$ | | 22.21 | 1 | 3.84 | 6.63 |

Model Formula 1 yields a good fit, whereas model Formula 4 yields a poor fit. Moreover, the difference of $G^2$-values of both models is significant at the one percent level. The obvious conclusion is that the item is uniformly biased. Table 3 shows that in all score categories Whites perform better than Blacks.

Table 3

Natural Logarithm Response Ratio and Parameter Estimates model Formula 1, Scheuneman's Data

| | | Black | White |
|---|---|---|---|
| | $\hat{G}_j$ | -.534 | .534 |
| Score | $\hat{S}_i$ | | |
| 13-14 | 1.526 | 1.992 | 2.996 |
| 12 | .708 | 1.099 | 2.197 |
| 10-11 | -.288 | -.083 | 1.314 |
| 1-9 | -1.946 | 1.293 | -.581 |

Note. The estimate of the constant C is .946.

The Rasch Model as a Loglinear Model

In the Rasch model for binary scored items the probability that subject $\underline{i}$ gives a response scored one to item $\underline{j}$ is written as (Rasch, 1960):

$$\underline{P}_{ij} = \exp(\underline{A}_i - \underline{D}_j)/\{1+ \exp(\underline{A}_i - \underline{D}_j)\}, \qquad (11)$$

where $\underline{A}_i$ represents subject's ability and $\underline{D}_j$ item's difficulty. From model Formula 11 follows that the logarithm of the response ratio is:

$$\ln\{\underline{P}_{ij}/(1 - \underline{P}_{ij})\} = \underline{A}_i - \underline{D}_j. \qquad (12)$$

Suppose a test is administered to a group of subjects. After removing all items that are scored zero or one by all subjects the test consists of $\underline{n}$ binary items. Under the Rasch model a subject's score, i.e. the number of one's, is the sufficient statistic for his or her ability (Fischer, 1974, p. 203). Therefore the estimates of the parameters can be obtained as if every subject with the same score has the same ability. The number of score groups is $(\underline{n} + 1)$, i.e. the scores run from 0 to $\underline{n}$. In general $\underline{p}$ of these score groups do not yield any information. This is always the case for the score groups 0 and $\underline{n}$ because all subjects in these groups have obtained a zero, respectively, one on all items. Moreover, it is possible that some score groups do not contain subjects at all. Therefore, only $(\underline{n} + 1 - \underline{p})$

score groups yield relevant information. The response variable has only two categories, zero or one. The information can be summarized in a $\underline{n} \times (\underline{n} + 1 - \underline{p}) \times 2$ Item x Score x Response table. As an example consider the frequencey table derived from Table 2 of Perline, Wright and Wainer's (1979) nine-item scale for parole decision data:

Table 4

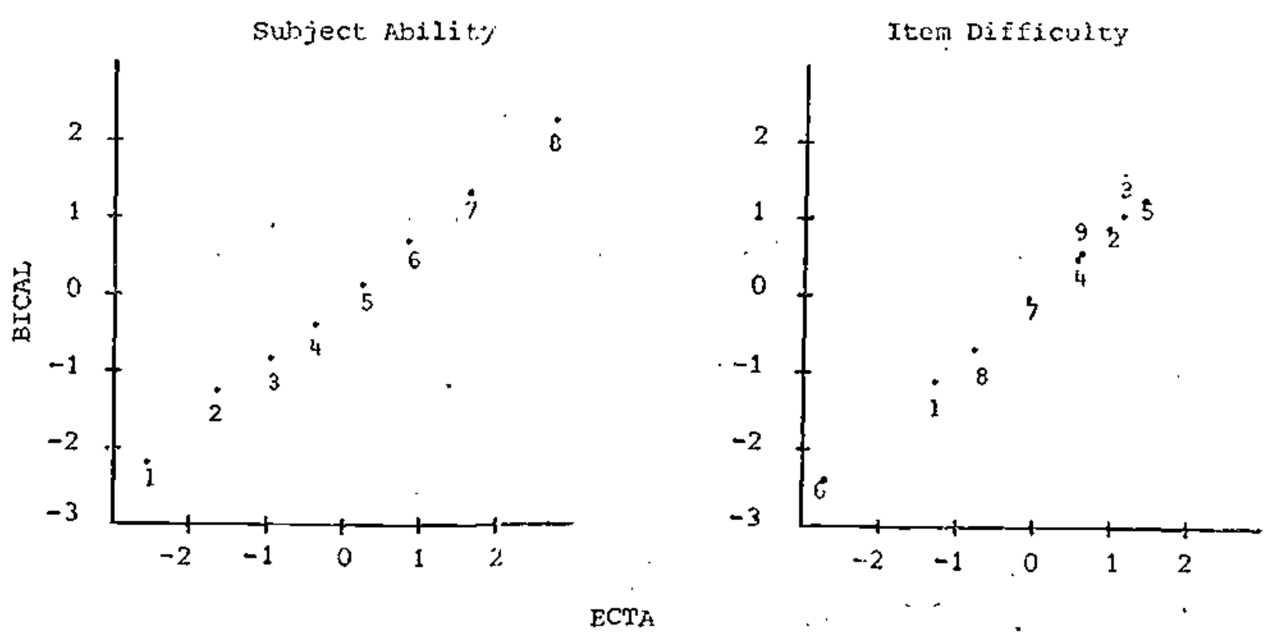
Frequencies in Item x Score x Response Table for Parole Data
(Perline, Wright & Wainer, 1979)

| | Response (m) | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item (j) | 1 | | | | | | | | 0 | | | | | | | |
| | Score (i) | | | | | | | | Score (i) | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 6 | 0 | 3 | 4 | 15 | 11 | 11 | 10 | 8 | 15 | 44 | 57 | 69 | 71 | 75 | 50 | 39 |
| 1 | 0 | 2 | 9 | 20 | 27 | 24 | 28 | 40 | 15 | 45 | 52 | 64 | 55 | 62 | 32 | 7 |
| 8 | 0 | 2 | 5 | 10 | 25 | 55 | 51 | 47 | 15 | 45 | 56 | 74 | 57 | 31 | 9 | 0 |
| 7 | 0 | 9 | 24 | 34 | 42 | 50 | 49 | 46 | 15 | 38 | 37 | 50 | 40 | 36 | 11 | 1 |
| 4 | 0 | 3 | 11 | 44 | 60 | 82 | 60 | 47 | 15 | 44 | 50 | 40 | 22 | 4 | 0 | 0 |
| 9 | 0 | 11 | 20 | 43 | 56 | 78 | 56 | 47 | 15 | 36 | 41 | 41 | 26 | 8 | 4 | 0 |
| 2 | 4 | 24 | 37 | 54 | 56 | 66 | 54 | 47 | 11 | 23 | 24 | 30 | 26 | 20 | 6 | 0 |
| 3 | 0 | 10 | 32 | 57 | 69 | 83 | 58 | 47 | 15 | 37 | 29 | 27 | 13 | 3 | 2 | 0 |
| 5 | 11 | 30 | 41 | 59 | 64 | 67 | 54 | 47 | 4 | 17 | 20 | 25 | 18 | 19 | 6 | 0 |

Comparing the linear logit model Formula 1 for this table with the Rasch model Formula 12 shows that the models are equivalent. Consequently, the loglinear model can be used for estimating the parameters and the goodness of fit of the Rasch model. Using the program ECTA the model Formula 1 was fitted to Perline, Wright and Wainer's data. The fit of the model is rather poor: .The values of Pearson's and likelihood ratio chi-squares are, respectively, 285.54 and 295.56 with 56 degrees of freedom. The parameter estimates from the ECTA program were compared to the BICAL program Rasch model estimates reported by Perline, Wright, and Wainer. As Figure 1 shows the relation of the ECTA and BICAL estimates is almost perfectly linear.

Figure 1

Plot of Estimates Computed from BICAL (Perline, Wright & Wainer, 1979) versus ECTA for Parole Data



In the loglinear model the parameters are estimated using the maximum likelihood method. The question is whether the estimates are unconditional (UML) or conditional maximum likelihood (CML) estimates.

CML estimates are obtained maximizing the likelihood function with respect to a parameter vector $\underset{\sim}{\theta}_1$ conditional on the minimal sufficient statistics for the remaining model parameters in the vector $\underset{\sim}{\theta}_2$ (Andersen, 1973, p. 37). The estimate for the vector $\underset{\sim}{\theta}_1$ is a CML estimate with respect to $\underset{\sim}{\theta}_2$. But the estimates for the parameters in $\underset{\sim}{\theta}_1$ are UML estimates with respect to each other.

Using the linear logit model Formula 1 implies that the Item x Score table is considered to be fixed by the design. The product multinomial sampling is the only design in which conditioning on sufficient statistics is used (Bishop, Fienberg & Holland, 1975, p. 63). The likelihood function is conditional on the sufficient statistics for the parameters of the Item x Score table, which are fixed by the design. However, this function contains all the parameters of Formula 1. The likelihood is maximized simultaneously with respect to the parameters $\underline{C}$, $\underline{S}_i$ and $\underline{G}_j$. Consequently the estimates for the parameters in the log-linear model yield UML estimates for the parameters in the Rasch model.

Acknowledgement

The authors thank Fred N. Kerlinger for his comments.

## References

Andersen, E.B. Conditional inference and models for measuring. Copenhagen, Denmark: Mentalhygienisk Forlag, 1973.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. Discrete multivariate analysis: Theory and practice. Cambridge, Mass.: MIT Press, 1975.

Dixon, W.J., & Brown, M.D. Biomedical computer programs, P-series 1977. Berkely: University of California Press, 1977.

Fienberg, S.E. The analysis of cross-classified data. Cambridge, Mass.: MIT Press, 1977.

Fischer, G. Einführung in die Theorie psychologischer Tests. Bern: Huber, 1974.

Goodman, L.A., & Fay, R. ECTA program, description for users. Chicago: Department of Statistics, University of Chicago, 1974.

Perline, R., Wright, B.D., & Wainer, H. The Rasch model as additive conjoint measurement. Applied Psychological Measurement, 1979, 3, 237-255.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: The Danisch Institute for Educational Research, 1960.

Scheuneman, J. A method of assessing item bias. Journal of Educational Measurement, 1979, 16, 143-152.